Health Status: Types of Validity and the Index of Well-being

by Robert M. Kaplan, J.W. Bush, and Charles C. Berry

The concept of validity as it applies to measures of health and health status is examined in the context of a set of standard, widely accepted definitions of validity. Criterion validity is shown to be irrelevant to health status measures because of the lack of a single specific, directly observable measure of health for use as a criterion. To overcome this problem, the Index of Well-being has been constructed to fulfill the definition of content *validity by including all levels of function and symptom/problem complexes, a clearly defined relation to the death state, and consumer ratings of the relative desirability of the function levels. Data from a two-wave household interview survey provide convergent evidence of* construct *validity by demonstrating an expected positive correlation of the Index of Well-being with self-rated well-being and expected negative correlations with age, number of chronic medical conditions, number of reported symptoms or problems, number of physician contacts, and dysfunctional status. Discriminant evidence of* construct *validity is demonstrated by predicted differences in correlation between concurrent Index of Well-being scores and self-assessed overall health status, and between the Index of Well-being scores and self-rated well-being on different days. A simple method of estimating a currently usable comprehensive population index of health status, the Weighted Life Expectancy, is described.*

Leaders, researchers, and decision makers in health services need a comprehensive numerical expression of health status for three distinct but related uses. First, empirical evaluation of health programs that care for diverse patient populations is impossible without a measure that aggregates different outcomes, including death, from many different health problems on a single scale [1]. Second, sensitive estimation of the probable effects of proposed new programs

Research supported by grant no. 2 R18 HS00702 from the National Center for Health Services Research, DHEW. An earlier version of this article was presented at a symposium, "Health Status Indexes: Their Role in Tomorrow," at the annual meeting of the American Association for the Advancement of Science, Boston, Feb. 1976.

Address communications and requests for reprints to J.W. Bush, M.D., Director, Division of Health Policy C-007, Department of Community Medicine, University of California at San Diego, La Jolla, CA 92093. Robert M. Kaplan, Ph.D., is assistant professor of psychology at San Diego State University. Charles C. Berry, Ph.D., is assistant professor of community medicine at the University of California at San Diego.

requires a common effectiveness measure for policy analysis and resource allocation [2]. Third, comparing the health status of different populations at different times requires a single social indicator for health, since the results of present comparisons now depend on which indicator is selected [3,4].

A number of attributes have been suggested for an ideal health status index [5–11]. Our research group has proposed a series of closely related indexes, which have been designed to possess most if not all of these desirable attributes and which have potential for serving all three of the above-mentioned uses; these include the Weighted Life Expectancy, which is derived from the Index of Well-being described in several publications [1–3,12–20]. One of the most important attributes, yet one of the most complex and confusing, is validity. The purpose of this article is twofold: first, to clarify the meaning of the term "validity" as it applies to health status measures in general, and, second, to present a preliminary assessment of the validity of the Index of Well-being (IWB), the time-specific component of a comprehensive measure of health status, that is, the Weighted Life Expectancy.

Validity and Health Status

The subject of validity is a frequently misunderstood problem in health status measurement. Therefore health index researchers (and their critics) are obligated to be as clear, explicit, precise, and consistent as possible in the terms they use to describe the data and relationships that they offer (or will accept) as evidence for validity.

Validity indicates the range of inferences that are appropriate when interpreting a measurement, a score, or the result of a test [21]. That is, the validity of a measure defines the meaning of a score. Validity is not absolute; it is relative to the domain about which statements are made. If we want to measure what society means by "health," then an indicator or index is a valid measure of total health status only to the extent that it expresses or quantifies that construct.

Because of the uncertainties, judgments, and assumptions that are required in assessing the correspondence between an operational measurement and a conceptual variable, different researchers have proposed many different methods —labeled with many different names—to assess validity. To help minimize confusion, a joint committee of the American Psychological Association, the American Education Research Association, and the National Council on Measurement in Education, in their *Standards for Educational and Psychological Tests* [22], defined three basic types of validity: criterion, content, and construct.

These three types subsume almost all the forms of validity that have been proposed. "Concurrent" validity and "predictive" validity, for example, are subcategories of *criterion* validity. "Empirical" validity and "statistical" validity are synonyms for criterion validity. "Convergent" validity and "discriminant" validity [23] are really types of evidence for *construct* validity. "Trait" validity

and "factorial" validity are sometimes considered synonyms for construct validity [24]. "Face" validity is not considered a form of validity at all, whereas *content* validity is strictly defined [22]. Final acceptance of the validity of a measure or a theory depends on the collective judgments of persons knowledgeable in a field. In the absence of consistent discernible criteria for that acceptance, the acceptance itself has been called "consensual" validity [25].

Criterion Validity

A proposed measure achieves criterion validity (or empirical or statistical validity) to the extent that it corresponds to some other observation that measures *accurately* the phenomenon of interest. If the proposed measure corresponds to a criterion measured simultaneously (for example, as blood pressure cuff measurements correspond to intra-arterial pressure measured at the same time), the validity is called "concurrent." If the proposed measure forecasts a future criterion value (as a present test score predicts future job performance), the validity is called "predictive."

By definition, the criterion must be a *superior*, more accurate measure of the phenomenon if it is to serve as a verifying norm. If a criterion exists, only greater practicality or less expense justifies the use of concurrent measures as proxies. If the criterion is not a superior measure, then failure of correspondence by any new measure may be a defect of the criterion itself, making it insufficient as a reference for validity.

The standard National Health Interview Survey (NHIS) list of chronic medical conditions, for example, has been referred to as a criterion, apparently concurrent, for health measurement [26]. That list, however, does not include any acute, transient conditions and does not measure current function at all. Furthermore, since many people have a very poor understanding of their diagnoses (or even of the word "chronic"), the list does not provide any medical verification of an index. Indeed, if the list of chronic conditions could be taken as a criterion, it should be used as the primary measure, since the list takes even less time than the 5–10 minutes required for a full battery of function status questions.

Since few contend that the number of chronic medical conditions accurately represents society's notion of health status, referring to the NHIS list or to some similar single count as a "criterion" confuses the issue of validity. Such indicators can, however, provide useful convergent evidence for construct validity, as we shall discuss later.

The lack of a criterion—a single measure that corresponds even roughly to what is meant by "well-being" or "health status"—is the first and foremost reason for developing an index. As a social construct, "health" aggregates multiple observations on several dimensions across the total spectrum of dysfunctions that diseases and injuries impose on all members of the population. Accurate expression of the total concept requires a derived measure or index number—a combination of many different, fundamental, directly observed measures. Because no well-defined criteria exist for health as a social phe-

nomenon, the question of validity must be approached in broader terms than those traditionally used to assess medical and morbidity measures. This difference in approach complicates the assessment of validity [27].

Predictive Validity. Predictive validity requires prior selection of an outcome criterion, just as concurrent validity does; for predictive validity a future value of that criterion must be matched. Diagnoses, laboratory tests, and physiological and other medical data implicitly involve forecasting and thus the estimation of prognoses. The validity of the prognostic components in a health index is not a difficult or controversial conceptual problem. Prognoses are directly analogous to the probabilities in the life table. Although statistical estimation is complex, the major problem is definition and aggregation of the multiple states among which movement must be represented. When a comprehensive set of states is available, the probability of movements among the states can be estimated—but prognostic information from different data sources cannot be compared until a comprehensive, disease-independent set of states is defined.

Content Validity

Content validity depends on whether the items of an instrument adequately represent the domain they are supposed to measure. The test construction procedures must indicate that all dimensions generally considered relevant have helped to define the domain and that the domain was appropriately sampled.

A measure with content validity will almost certainly exhibit "face validity," which is the simple appearance that the items are related to the construct of interest. However, the joint committee, in contrast to many authors, subtly but sharply distinguishes between content validity, which is legitimate, precisely defined, and necessary, and face validity, which is not an appropriate or reliable basis for inference [22].

The theoretical construct chosen for an index of health status guides the selection of the content, and the content in turn provides support for the construct. The recommendation of the joint committee [22] is that content should be viewed as an independent form of (internal) evidence to complement convergent and discriminant (external) evidence supporting validity.

Construct Validity

Before 1950, most social scientists considered only criterion and content forms of validity. By the mid 1950s, investigators had concluded that no clear criteria existed for most of the social measures being developed [24]. Developing a measure of intelligence, for example, was difficult since a precise definition of intelligence was lacking. The evolution that occurred was to establish more explicit foundations for the assessment of construct validity.

Construct validity involves positively specifying the dimensions of the construct, the domain of the dimensions both uniquely and jointly, and the expected relations of the dimensions to each other, both internally and externally. This process is required when "no criterion or universe of content is

accepted as entirely adequate to define the quality to be measured" [28]. Construct validation involves assembling empirical evidence to support the inference that a particular measure has meaning. It is an ongoing process, akin to amassing support for a complex scientific theory for which no single set of observations provides crucial or critical evidence.

The status of external evidence differs under the construct formulation from what it is when such evidence is used as a norm to establish criterion validity. Specifically, evidence that might be used to argue for criterion validity now becomes *convergent* evidence for the validity of the construct. Now, instead of seeking a perfect correlation with a putative criterion, one seeks only that direction and level of correlation between a single existing measure and the proposed measure that is suggested by knowledge of the construct. *Discriminant* evidence, on the other hand, indicates that the proposed measure correlates better with a second measure accepted as more closely related to the construct than it does with a third, more distantly related, measure.

To be valid, a construct must have valid *content*—that is, the content must positively and exhaustively define the dimensions of the construct and its measures. Since the full universe of content in the term "health" is not yet generally agreed on, final resolution of the definitional problems can come only through consensual validation of some proposed construct by researchers and other users.

Factorial Evidence. Factor analysis is so closely associated in psychology with establishing *construct* ("trait" or "factorial") validity that we must comment on it in detail, because we consider it generally inappropriate for constructing health indexes. In factor analysis (and its close relative, principal components analysis), the data ordinarily consist of the frequency distribution of the subjects over the items or variables in the set. The different items are all considered equally important (or unimportant), and correlations among the responses on each item are determined across all the subjects. In the usual application, the principal components of the correlation matrix yield factors, and the loading of a particular item on a particular factor is its "weight" on that factor. The percentage of the total variation explained by a given factor and its items is then usually referred to as the relative "importance" of the factor. Factors and items that contribute little to explaining variance in occurrence or frequency are considered "not to discriminate" and to be "unimportant." In general, such a procedure subtly substitutes variation in frequency for variation in social importance. Items that are checked rarely or are poorly correlated with other items in a given population may receive very low weights on all factors, or may even yield an independent factor with a very low eigenvalue, regardless of how important they may be. This may occur in a household survey for very low levels of function. Low levels of function are not frequent but are very important when they do occur, so such items should be retained.

The above problem is quite separate from problems caused by the fact that the correlational structure varies significantly among different patient and population groups. No universal population is definable for a general purpose

index since, in surveys of probability samples, many aspects of dysfunction occur so rarely that they cannot be correlated reliably.

In addition, the variables in a factor analysis usually include uncontrollable as well as controllable sets. A health program does not generally change the correlations among uncontrollable variables, such as age, income, and education, but it attacks directly the occurrence of and the correlations among symptoms, problems, and other evidences of dysfunction. Since the two sets of variables are intimately related in the factor structure, the health program that significantly affects the controllable set will fundamentally alter the factor structure, invalidate the factor equation carried over from prior analyses as an outcome measure, and seriously bias any estimate of change. Similar changes in factor structure could occur from other nonprogrammatic influences as well.

Furthermore, among the variables that change in response to a health program or other influences, there is no reason to believe that the changes occur at the same rate on all variables or even in the same direction (as, for example, morbidity and mortality). Such differential rates of change will also alter the factor structure, preventing its use as a consistent summary or description of health status change.

Most factor analyses in health have been based on frequency data [29,30]. The items, variables, or rates in such studies are well understood, and aggregating them into underlying "dimensions" adds little to their interpretation: the factors certainly cannot be regarded as etiologies or causes. The analyses only demonstrate that many medical, health service, and social phenomena are correlated to some extent within different patient and population groups; the relation of the factors themselves to health programs or to the concept of health status in general is difficult, if not impossible, to define.

Factor analysis might be applied where the respondents rate their relative preference for a set of items or attributes. The factors would then represent correlations in the preference structure independent of the frequency of an item's occurrence. The relative preferences for various conditions have no reason to vary with the relative frequency of those conditions. Our survey data, for instance, indicate that "pain or discomfort from sexual organs" and "trouble learning, remembering, or thinking clearly" occur with very different frequencies but that the social preference for each is low. On the other hand, "feeling tired and weak" occurs with about the same frequency as "hearing difficulty," but the preferences for the two symptoms are significantly different. If an index were constructed from factor analyses of preference data alone, small uncorrelated differences in item preferences might not load significantly on any of the larger factors and might not represent a substantial unique factor. Such items might be eliminated regardless of how much their high frequency might affect the overall well-being of the total population.

These problems do not alter the value of factor analysis as a statistical or data reduction technique, especially among highly correlated independent variables. They indicate, however, that factor analysis is inappropriate for constructing an outcome or dependent variable where relative frequency and

proportions of variance explained may be substituted for social preferences. Factor analysis cannot derive measures of relative importance from measures of relative frequency any more than it can derive measures of relative frequency from measures of relative importance. The two are conceptually and empirically independent, and factor analysis cannot go beyond whatever data are being analyzed. A comprehensive health status measure must rationally and explicitly reflect differences in both preferences ("weights") and frequency of occurrence, and each phenomenon must be measured directly. Since factor analysis does not offer a means of combining the two components, factorial validity has limited value in constructing or validating health indexes.

The Index of Well-being: Content Validity

The IWB is the time-specific facet of a comprehensive construct of health status that includes two distinct components: (1) level of well-being and (2) prognosis. The term "well-being" is chosen to indicate that the dimension expressed represents the total quality of life in regard to health. If other aspects of life—e.g., housing and income—were to be included in a total quality of life index, the well-being associated with a given income and with a given kind of housing would be the variable to be measured and incorporated into the index. The number of persons per room might constitute an index of housing, for example, but it would not reflect the relative desirability (satisfaction, utility) of housing except by virtue of implicitly assumed or consensually established levels of quality, value, or well-being associated with various amounts of housing space. Thus the levels of well-being for our proposed health status index are defined by the subjective preferences or weights that members of society associate with time-specific states and function levels.

Prognoses are the probabilities of transition among the function levels, governed over time by disease and other disorders; prognoses are essential to describe the concept of health implicit in our decision making [13]. The comprehensive index views health status as an expectation: a joint function of the levels of well-being (the weights of the states) and the expected duration of stay in each state, derived from the prognoses.

The comprehensive index may be expressed as

$$E = \sum_{j=1}^{J} W_j Y_j \qquad (1)$$

where E is the Weighted (quality-adjusted) Life Expectancy for a cohort or population in well-year equivalents

Y_j is the expected duration in each function level j, computed from the transition probabilities [3,20], $j = 1, \ldots, J$

W_j is the level of well-being, i.e., the social preference weight associated with each function level j

J is the total number of function levels in a given analysis

Since it is based on the expected duration of stay Y_j in each level, the Weighted Life Expectancy E cannot be observed and is defined only for populations, not for individuals.

The construction of the Index of Well-being will be described in the course of defining the function levels and considering the content and construct validity of the IWB. For the present, we note that W represents a *symptom-standardized* Index of Well-being and that W^* represents a more refined *symptom-specific* Index of Well-being. They are related time-specific measures that represent a health situation, excluding prognoses, for a single day or a short period, and their validity contributes to the validity of the comprehensive index.

Function Levels

The concept underlying the time-specific Index of Well-being was to define the universe of all possible situations between optimum function and death that might serve as a classification matrix and sample frame [31]. From an extensive specialty-by-specialty review of medical reference works, we listed all the ways, however minor, that diseases and injuries can affect a person's behavior and role performance, regardless of etiology.

By matching the disruptions in role performance and other activities with standard survey items, we created separate scales for mobility, physical activity, and social activity and defined all the steps that were perceptibly different from one another [14]. Survey instruments have been developed that will classify a person into one and only one step of each of the three scales. Several research groups have now classified over 10,000 respondents on over 35,000 different days using these instruments [26,32]. Of the 100 theoretically possible combinations of the scale steps, we have now observed 43, which we refer to as function levels; these are shown in Table 1. These include several combinations that we originally considered unlikely; as others are observed, they will also be added to the list. Open-ended questions at the end of the instrument have revealed no other significant functional limitations except the unlikely combinations mentioned.

Having intentionally developed at the outset an overly refined classification, especially in the extreme levels of dysfunction, we are confident that our instruments now distinguish all meaningful function levels. Consolidating steps on the scales according to value studies and other criteria will further streamline the instruments and yet preserve adequate discrimination over the range from completely well to death.

Early in 1974 we began a two-year panel survey of San Diego area households to determine the validity and reliability of our classifications into the function levels and to determine the social preference weights. A probability sample of 867 respondents was interviewed. Data were also gathered about a supplementary probability sample of 370 children and 89 dysfunctional persons identified in the sample households by a screening question. One year later we reinterviewed 80 percent of the respondents about themselves, their children, and the dysfunctional persons in their households. In both 1974 and 1975

Table 1. Function Levels: Combinations of Steps on Mobility, Physical Activity, and Social Activity Scales, with Associated Levels of Well-being (Social Preference Weights), W_j

Numbers in parentheses are step numbers on the three scales.

Function level number (j)	Scale			Level of well-being (W_j)
	Mobility	Physical activity	Social activity	
		NO SYMPTOM/PROBLEM COMPLEX		
L 43	Drove car and used bus or train without help (5)	Walked without physical problems (4)	Did work, school, or housework, and other activities (5)	1.000
		SYMPTOM/PROBLEM COMPLEX PRESENT		
L 42	Drove car and used bus or train without help (5)	Walked without physical problems (4)	Did work, school, or housework, and other activities (5)	0.7433
L 41	Drove car and used bus or train without help (5)	Walked without physical problems (4)	Did work, school, or housework, but other activities limited (4)	0.6855
L 40	Drove car and used bus or train without help (5)	Walked without physical problems (4)	Limited in amount or kind of work, school, or housework (3)	0.6683
L 39	Drove car and used bus or train without help (5)	Walked without physical problems (4)	Performed self-care, but not work, school, or housework (2)	0.6955
L 38	Drove car and used bus or train without help (5)	Walked without physical problems (4)	Had help with self-care activities (1)	0.6370
L 37	Drove car and used bus or train without help (5)	Walked with physical limitations (3)	Did work, school, or housework, and other activities (5)	0.6769
L 36	Drove car and used bus or train without help (5)	Walked with physical limitations (3)	Did work, school, or housework, but other activities limited (4)	0.6172
L 35	Drove car and used bus or train without help (5)	Walked with physical limitations (3)	Limited in amount or kind of work, school, or housework (3)	0.6020
L 34	Drove car and used bus or train without help (5)	Walked with physical limitations (3)	Performed self-care, but not work, school, or housework (2)	0.6292
L 33	Drove car and used bus or train without help (5)	Walked with physical limitations (3)	Had help with self-care activities (1)	0.5707
L 32	Did not drive, or had help to use bus or train (4)	Walked without physical problems (4)	Did work, school, or housework, but other activities limited (4)	0.6065

Table 1. Continued.

Numbers in parentheses are step numbers on the three scales.

Function level number (j)	Scale			Level of well-being (W_j)
	Mobility	Physical activity	Social activity	
L 31	Did not drive, or had help to use bus or train (4)	Walked without physical problems (4)	Limited in amount or kind of work, school, or housework (3)	0.5913
L 30	Did not drive, or had help to use bus or train (4)	Walked without physical problems (4)	Performed self-care, but not work, school, or housework (2)	0.6185
L 29	Did not drive, or had help to use bus or train (4)	Walked without physical problems (4)	Had help with self-care activities (1)	0.5600
L 28	Did not drive, or had help to use bus or train (4)	Walked with physical limitations (3)	Did work, school, or housework, but other activities limited (4)	0.5402
L 27	Did not drive, or had help to use bus or train (4)	Walked with physical limitations (3)	Limited in amount or kind of work, school, or housework (3)	0.5250
L 26	Did not drive, or had help to use bus or train (4)	Walked with physical limitations (3)	Performed self-care, but not work, school, or housework (2)	0.5523
L 25	Did not drive, or had help to use bus or train (4)	Moved own wheelchair without help (2)	Limited in amount or kind of work, school, or housework (3)	0.5376
L 24	Did not drive, or had help to use bus or train (4)	Moved own wheelchair without help (2)	Performed self-care, but not work, school, or housework (2)	0.5649
L 23	In house (3)	Walked without physical problems (4)	Performed self-care, but not work, school, or housework (2)	0.6488
L 22	In house (3)	Walked without physical problems (4)	Had help with self-care activities (1)	0.5902
L 21	In house (3)	Walked with physical limitations (3)	Did work, school, or housework, but other activities limited (4)	0.5704
L 20	In house (3)	Walked with physical limitations (3)	Limited in amount or kind of work, school, or housework (3)	0.5552
L 19	In house (3)	Walked with physical limitations (3)	Performed self-care, but not work, school, or housework (2)	0.5824
L 18	In house (3)	Walked with physical limitations (3)	Had help with self-care activities (1)	0.5239
L 17	In house (3)	Moved own wheelchair without help (2)	Performed self-care, but not work, school, or housework (2)	0.5950

Table 1. Continued.

Numbers in parentheses are step numbers on the three scales.

Function level number (j)	Scale			Level of well-being (W_j)
	Mobility	Physical activity	Social activity	
L 16	In house (3)	Moved own wheelchair without help (2)	Had help with self-care activities (1)	0.5364
L 15	In house (3)	In bed or chair (1)	Performed self-care, but not work, school, or housework (2)	0.5715
L 14	In house (3)	In bed or chair (1)	Had help with self-care activities (1)	0.5129
L 13	In hospital (2)	Walked without physical problems (4)	Performed self-care, but not work, school, or housework (2)	0.6057
L 12	In hospital (2)	Walked without physical problems (4)	Had help with self-care activities (1)	0.5471
L 11	In hospital (2)	Walked with physical limitations (3)	Performed self-care, but not work, school, or housework (2)	0.5394
L 10	In hospital (2)	Walked with physical limitations (3)	Had help with self-care activities (1)	0.4808
L 9	In hospital (2)	Moved own wheelchair without help (2)	Performed self-care, but not work, school, or housework (2)	0.5520
L 8	In hospital (2)	Moved own wheelchair without help (2)	Had help with self-care activities (1)	0.4934
L 7	In hospital (2)	In bed or chair (1)	Performed self-care, but not work, school, or housework (2)	0.5284
L 6	In hospital (2)	In bed or chair (1)	Had help with self-care activities (1)	0.4699
L 5	In special care unit (1)	Walked without physical problems (4)	Performed self-care, but not work, school, or housework (2)	0.5732
L 4	In special care unit (1)	Walked without physical problems (4)	Had help with self-care activities (1)	0.5147
L 3	In special care unit (1)	Walked with physical limitations (3)	Performed self-care, but not work, school, or housework (2)	0.5070
L 2	In special care unit (1)	Walked with physical limitations (3)	Had help with self-care activities (1)	0.4483
L 1	In special care unit (1)	In bed or chair (1)	Had help with self-care activities (1)	0.4374
L 0	Dead (0)	Dead (0)	Dead (0)	0.0000

we gathered data for each person on sociodemographic characteristics, role performance, number and kinds of symptoms and problems, number of physician contacts, number of chronic conditions, and self-ratings of current well-being and overall health status. The respondents also gave category ratings for a set of case descriptions which we used to compute social preferences for a series of function levels and symptom/problem complexes.

To investigate the validity and reliability of our instruments in classifying respondents into one and only one of the function levels, we are using tape recordings of the interviews and other information to analyze a randomly counterbalanced field experiment that was built into our 1974 survey, comparing a show-card and a branching-question form of the instruments. The major result of these comparisons has been to reveal a possible cause of widespread under-reporting in health surveys: the phrasing of questions in terms of self-assessed capacity ("could," "able," "needed," "required," and the like) rather than in behavioral terms ("did"). We have purged our instruments of all questions that elicit what is really a judgment or opinion about capacity, and our function level classifications are now based strictly on behavioral criteria.

This focus on behavior makes it possible to classify dysfunctions from acute and chronic illness and from disability clearly on the same scale and to record daily and short-term changes in the function levels. Incorporating both transient and long-term disturbances on the same scale is an important aspect of content validity for a health index.

Symptom/Problem Complexes

Using a procedure analogous to construction of the function levels, we exhaustively listed all possible symptoms and problems and aggregated them into frequently occurring groups, as shown in Table 2. In response to a list, the respondents report all the symptom/problem complexes that they experienced on a given day, without attaching any rating of severity. In a follow-up question, the respondent then selects the symptom or problem experienced as the "most undesirable" on that day.

In the 1974 survey, an open-ended follow-up question revealed approximately 15 variants of symptoms and problems that were not explicitly included in the initial set. The basic list and the new causes of disturbed role performance were then consolidated, according to medical and value criteria, into our current list (Table 2), which now numbers 36. No symptoms or problems were eliminated because of arbitrary statistical criteria: all problems, however small, are included in the weighting scheme in combination with the function level information. The most common symptom unaccounted for in 1974 was depression or anxiety, so a new complex was added to the 1975 list: "Spells of feeling upset, depressed, or crying." We plan eventually to use this flexible approach to give more detailed coverage to other emotional disturbances; this approach also permits the systematic addition of new aspects of health status as they are defined [33].

The method of test construction and the results of multiple surveys by our-

Table 2. Symptom/Problem Complexes and Linear Adjustments W_i for Symptom-specific Level of Well-being Scores

Complex no. (i)	Symptom/problem complex	Adjustment (W_i)
C 1	Any trouble seeing—includes *wearing glasses* or *contact lenses*	0.0190
C 2	Pain or discomfort in one or both eyes, such as burning or itching	0.0337
C 3	Trouble hearing—includes wearing *hearing aid*	0.0834
C 4	Earache, toothache, or pain in jaw	0.0978
C 5	Sore throat, lips, tongue, gums, or stuffy, runny nose	0.0933
C 6	Several or all permanent teeth missing or crooked—includes wearing bridges or dentures (false teeth)	0.0715
C 7	Pain, bleeding, itching, or discharge (drainage) from sexual organs—does *not* include normal menstrual (monthly) bleeding	−0.0920
C 8	Itching, bleeding, or pain in rectum	−0.0379
C 9	Pain in chest, stomach, side, back, or hips	−0.0382
C 10	Cough *and* fever or chills	0.0077
C 11	Cough, wheezing, or shortness of breath	−0.0075
C 12	Sick or upset stomach, vomiting, or diarrhea (watery bowel movements)	0.0065
C 13	Fever or chills with aching all over *and* vomiting or diarrhea (watery bowel movements)	−0.0722
C 14	Hernia or rupture of abdomen (stomach)	−0.0501
C 15	Painful, burning, or frequent urination (passing water)	−0.0327
C 16	Headache, dizziness, or ringing in ears	0.0131
C 17	Spells of feeling hot, nervous, or shaky	0.0129
C 18	Weak or deformed (crooked) back	−0.0474
C 19	Pain, stiffness, numbness, or discomfort of neck, hands, feet, arms, legs, ankles, or several joints together	−0.0344
C 20	One *arm and one leg* deformed (crooked), paralyzed (unable to move), or broken—includes wearing artificial limbs or braces	−0.0681
C 21	One *hand or arm* missing, deformed (crooked), paralyzed (unable to move), or broken—includes wearing artificial limbs or braces	−0.0609
C 22	One *foot or leg* missing, deformed (crooked), paralyzed (unable to move), or broken—includes wearing artificial limbs or braces	−0.0630
C 23	*Two legs* deformed (crooked), paralyzed (unable to move), or broken—includes wearing artificial limbs or braces	−0.0881
C 24	*Two legs* missing—includes wearing artificial limbs or braces	−0.1027
C 25	Skin defect of face, body, arms, or legs, such as scars, pimples, warts, bruises or changes in color	0.0633
C 26	Burning or itching rash on large areas of face, body, arms, or legs	0.0171
C 27	Burn over large areas of face, body, arms, or legs	−0.1100
C 28	Overweight for age and height	0.0785
C 29	General tiredness, weakness, or weight loss	−0.0027
C 30	Trouble talking, such as lisp, stuttering, hoarseness, or being unable to speak	0.0194
C 31	Trouble learning, remembering, or thinking clearly	−0.0830
C 32	Loss of consciousness such as seizures (fits), fainting, or coma (out cold or knocked out)	−0.1507
C 33	Taking medication or staying on prescribed diet for health reasons	0.1124
C 34	Breathing smog or unpleasant air	0.1555
C 35	No symptom or problem	0.2567
C 36	Spells of feeling upset, depressed, or crying	. . .

selves and others support the conclusion that we have summarized a full array of the symptoms and problems, except those of mental health, that occur in people's daily lives. The exhaustiveness and simplicity of the symptom/problem classification help assure the content validity of the index.

The symptom/problem complex list is the major source of differentiation among persons who are completely functional, that is, who occupy steps (5,4,5) on the three scales of mobility, physical activity, and social activity. In fact, the basis for distinguishing level L42 from level L43 in our classification is the presence of symptoms and problems without any disturbance of function. In our 1974 survey, for example, among the 1,236 adults and children in the probability sample, 32.4 percent were in function level L43—steps (5,4,5) with no symptoms or problems; 48.9 percent were in level L42—steps (5,4,5) with at least one symptom or problem present; and 18.7 percent were distributed among all lower levels. Only persons in level L43 receive an index score of 1.0. Any symptom, however minor, significantly decreases the Index of Well-being.

This sensitivity of the index, advocated from our earliest proposals [12], is designed to describe small gradations in the upper levels of well-being and is an important part of the overall content validity of the index. Unfortunately, some other groups of researchers working with these function level scales have chosen not to include symptoms and therefore not to distinguish between the two top levels in their work. Not accounting for symptoms is a major sacrifice of content validity that we strongly discourage.

At the other extreme of the scale, both the function levels and their weights bear a clearly defined relation to the death state, which is assigned a value of zero. The importance of death makes its relation to all other aspects of health status one of the most important components of even the time-specific Index of Well-being. Any outcome measure that omits death from the analytical framework, and then omits deaths from the set of observations, will almost invariably observe the usual paradoxical relation between morbidity and mortality measures: the remaining living members of a population or patient group will show a higher mean level of well-being if death is omitted. This paradoxical relation necessarily biases all nonmortality-based estimates of health status change and program effectiveness, to favor those groups in which the most dysfunctional members are permitted to die [3]. To achieve content validity, therefore, even time-specific aspects of health status must have a clearly defined relation to mortality.

Social Preference Weights

Consumers do not regard the function levels and symptom/problem complexes as equally important or undesirable. To achieve content validity, therefore, we must incorporate the affective aspect of reported dysfunctions into the overall index. That is, we want to locate the function levels and symptom/problem categories on an interval (ideally ratio) scale of relative well-being.

It is important to note that the numbers attached to the function levels (0–43), the scale steps (0–5,0–4,0–5), and the symptom/problem complexes (1–36) are only labels. Although these scale step labels *may* be ordinal, there is little reason to believe that they possess interval properties. There is no theoretical or empirical justification, therefore, for performing arithmetic operations on the scale step labels as some have suggested. Such operations substitute

statistical assumptions for measures of social importance. A measure that distinguishes "better" from "worse" is impossible to create without using weights, at least implicitly. Use of *any* scales of dysfunction without measures of relative importance omits a critical element of content validity and introduces substantial bias by assuming equal weights among the items.

To derive measured weights or social preferences for the levels and symptom/problem complexes, we asked respondents in our 1974 and 1975 surveys to rate a series of case descriptions selected from a matrix bounded by the 43 function levels and the 36 symptom/problem complexes, as described in our previous studies [14,15]. The case descriptions simultaneously presented age, function level, and symptom/problem information to the respondent. Such composite ratings eliminate the need to assume additivity across all the attributes to derive single weights for actually observed cases.

Case descriptions for rating were selected in a series of balanced designs to allow sensitive tests for interactions and to maximize the power of a predictive statistical model by which a precise weight could be assigned to any of the 1,548 (43×36) theoretically possible cases. These preference ratings constitute the weights that reflect the social values or relative importance that society attaches to the various function levels. They are shown in the last column of Table 1.

In other research we have investigated the properties of the preference ratings and have established that:

1. Preferences can be measured reliably ($r = 0.91$) from cross-validation studies using randomly created parallel forms of the procedure [14], even without using questionable shrinkage adjustments for possible unreliability in the measurement procedure [34].
2. The values on the 0–1 scale possess equal-interval properties [15,16].
3. The category ratings are stable across different orders of testing and modes of test administration [15].
4. Linear statistical models accurately represent and predict ($R^2 \geqslant 0.96$) the mean and median global consumer ratings for individual case descriptions [17].
5. Age groups representing different phases of the life cycle account for only about 1 percent of the variance in the preference ratings [17].
6. Category ratings are consistent with results from magnitude estimation methods that produce ratio scales with an origin at death [15,18].
7. The preferences are generalizable across different social groups and their leaders, all of whom seem to share a consensus on the terminal values associated with the function levels [15].
8. The category ratings are consistent with results from axiomatically derived procedures, like the Von Neumann and Morgenstern standard gamble, that imply social choice [15,19].

With data now available, we will soon be able to examine the stability of the preferences over time.

Structure of the IWB

Given the foregoing definitions of the function levels and social preference weights, we can express W, the mean time-specific (cross-sectional) *symptom-standardized* Index of Well-being for a population, as a simple weighted average:

$$W = \frac{1}{N}\sum_{j=1}^{J} W_j N_j \qquad (2)$$

where N is the total number of persons in a population
N_j is the number of persons in each function level j, $j = 1, \ldots, J$
W_j is the social preference weight for each function level j, $j = 1, \ldots, J$
J is the total number of function levels

In this model the assigned weight is standardized by means of a linear statistical model of the preference measurements to adjust for the presence of symptoms and problems in each function level. The standardization includes level L42 (steps 5,4,5), where a standardized weight of 0.7433 (Table 1) is assigned if any symptom or problems at all are present.

The expression given above for W is a population index; for an individual the symptom-standardized level of well-being is simply that value of W_j that pertains to his observed function level. We can further refine the individual's score by adjusting the weight assigned to his function level by the weight of the symptom or problem (shown in the last column of Table 2) that he experienced and reported as "most undesirable" for each day. This adjustment produces the *symptom-specific* Index of Well-being, W^*_k: $W^*_k = W_j + W_i$, where W_j is the standardized social preference weight for function level j, as shown in Table 1, and W_i is the weight assigned to each specific symptom/problem complex i, shown in Table 2 (p. 490). As an example, a person in function level L30 who reports symptom/problem complex C10, "cough and fever or chills," would receive a score $W^*_k = 0.6185 + 0.0077 = 0.6262$. A completely well person's score (function level L43) is adjusted by the weight of the dummy complex C35: $0.7433 + 0.2567 = 1.0$.

As with the symptom-standardized index W, the population index for the *symptom-specific* Index of Well-being, W^*, is computed as a weighted average:

$$W^* = \frac{1}{N}\sum_{k=1}^{K} W^*_k N_k \qquad (3)$$

where k is an index for a particular case type, or combination of i and j, i.e., of function level (L_j) and symptom/problem complex (C_i), $k = 1, \ldots, K$
K is the total number of different types of cases in the analysis
W^*_k is the symptom-specific level of well-being assigned to a person of case type k
N_k is the number of persons of each case type k
N is the total number of persons in the population

Both methods of computing the index take account of symptom/problem complexes even in the topmost function level; both yield mean time-specific cross-sectional indexes.

"Subjective" ratings of various states of functioning are thus included in the index as a separately measured component, conceptually and mathematically independent of the particular respondent. The composite weights W_j and W_i are derived from the ratings of many cases by large numbers of respondents. Any desired degree of precision for the weights may be obtained simply by increasing the sample size. The preference weights, therefore, are invariant across all applications of the index and do not contribute at all to error variance in the index used as a covariate or outcome measure. The method of index construction confers the invariance property, whether the weights are from the same population or another.

The construction of the IWB thus bypasses the problems of direct individual self-ratings. The function level classification is based on reports about observable behaviors on recent specific days, uncontaminated by personal or professional judgments about ability or need. Even when symptoms and problems are present, the respondent notes only their presence or absence and identifies the one considered "most undesirable."

Some investigators have suggested that the IWB is too "objective" because it does not give enough weight to individual feeling [26,35]. It is clear, however, that we have included an affective or preference component in the versions of the index that we recommend. Personal reports must inevitably be aggregated, for both statistical analysis and resource allocation; aggregating and standardizing the preference weights in advance avoids the extraneous variation inherent in self-ratings. Thus all changes or differences in the computed index relate only to differences in objectively reportable conditions. All persons who report the same function level for W_j (and "most undesirable" complex for W^*_k) receive exactly the same score.

In previous reports we have referred to the Index of Well-being as the Function Status Index. That title now seems undesirable for two reasons: it does not signify incorporation of the relative social desirability ("subjective aspect") of function status, and the Function Status Index title has recently been used [26] to refer to a count of the separate scales of mobility, physical activity, and social activity plus self-care on which a person registers no dysfunction. This count is, in effect, an implicit equal-weighting scheme for each scale that considers neither measured consumer preferences nor the death state and so departs significantly from the index that we advocate.

Content validity is enhanced when one knows what a measure does *not* contain. Because it does not incorporate prognoses, the Index of Well-being does not confuse the "expected future" with the present, as do indicators that do not clearly separate prognoses. Therefore the IWB can be used to monitor changes in health status over time, even though it is not by itself a complete indicator of health status but a component of a comprehensive index that does include prognoses.

In summary, the Index of Well-being excludes prognoses and includes symptom/problem complexes and exhaustive scales for mobility, physical activity, and social activity, with standardized consumer preference weights that apply across both acute and chronic dysfunctions. A simple weighted average explicitly relates all these aspects of function to the death state. Thus the IWB contains the elements and dimensions required to represent the time-specific aspects of a more comprehensive construct of health status that also includes prognoses. We believe, therefore, that the IWB fulfills the requirements of content validity discussed earlier.

The Index of Well-being: Construct Validity

Factor analysis is not useful in the construction of an indicator of health status like the IWB because of the statistical and conceptual problems described earlier. A health index of the type we are discussing must be constructed on the basis of substantive theory. The theoretical basis of the Weighted Life Expectancy is the concept of health status as an expectation, of which the IWB is the time-specific component. This concept is consistent with the standard life table and with theory from several disciplines.

The basic paradigm is from decision theory [36–38], with social preferences corresponding to utilities and the function levels corresponding to states, among which the system (or person) moves over time according to the prognoses, which correspond to transition probabilities [20,39]. The relationships explicit in the Index of Well-being and the Weighted Life Expectancy are consistent with the idea of illness as deviance from social norms [40], a general theory of disease [41], human information processing [42], and microeconomic theory, which now recognizes that consumers maximize their "stock of health," which is tantamount to the Weighted Life Expectancy [43,44]. Rather than using ad hoc statistical analyses, we have tried to formulate a model that rigorously relates all aspects of the health measurement problem and integrates them with relevant theory from contemporary social, management science, and medical disciplines.

Although they should not be considered operational definitions, constructs do imply empirical properties for proposed measures. Through a network of observations, the IWB can be demonstrated to be consistent with predictions made from the underlying concept. The construction of the index is an attempt to close the gap between a theoretical concept and its operational measurement. Thus the index reflects the empirical properties of the construct [45].

We contend that our proposed Index of Well-being contains almost all the time-specific content of a comprehensive health status measure (and, we hope, little else). We must now see if the data yielded by the IWB relate as expected to data yielded by other measures. Such relationships provide the two major types of external evidence for construct validity, *convergent* evidence and *discriminant* evidence. Because of their importance, these two types of evidence are frequently referred to as convergent validity and discriminant validity.

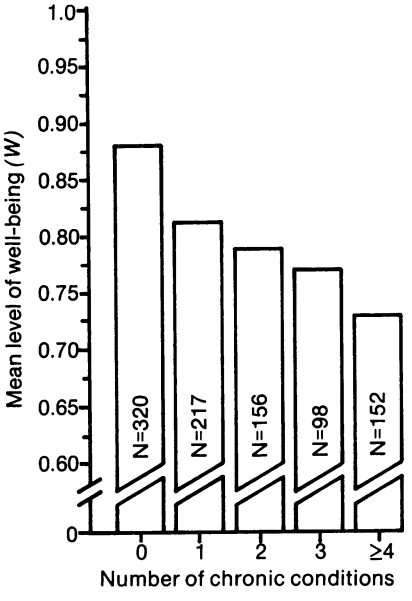

Fig. 1. Mean symptom-standardized level of well-being (W) for groups of persons reporting specific numbers of chronic conditions (weighted Pearson's $r = -0.96$).

Convergent Evidence

Convergent evidence is obtained either by showing that a test is related to other measures of the same phenomenon or by observing empirical relationships that can be predicted from the theoretical description of the construct. Establishing convergent validity for the IWB requires that it exhibit expected correlations, positive or negative, with other single variables accepted as relevant to or associated with well-being.

To test predictions about the relation of the Index of Well-being to other pertinent variables, we performed a variety of analyses on data from our San Diego surveys. These analyses represent only the initial studies of our data; in some cases more refined analyses will be possible. We present here only those correlations that would be accepted as evidence against validity if they did not have the expected sign. Because of the sample size, all of the correlations are statistically significant ($p < 0.001$).

Number of Chronic Conditions. Since chronic medical conditions produce both discomfort and disability, we would expect that a larger number of chronic conditions would, on the average, cause a lower level of well-being. Figure 1 displays the mean level of well-being for groups distinguished by number of chronic conditions. The mean level of well-being decreases monotonically as

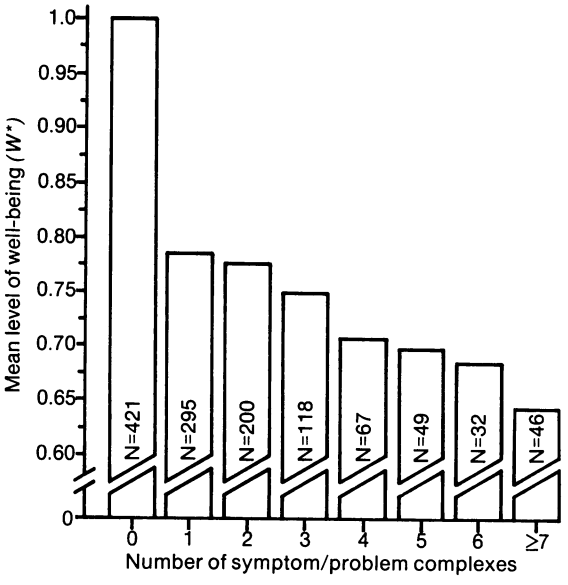

Fig. 2. Mean symptom-specific level of well-being (W*) for groups of persons reporting specific numbers of symptom/problem complexes (weighted Pearson's $r = -0.75$).

the number of chronic conditions increases; this relationship is consistent with observations by others [26].

Since the predicted relationship specifies only an aggregate decline with the number of chronic conditions, the appropriate statistic to express this relation is the correlation coefficient on the mean IWB weighted by the number of persons in each group. For the relation in Fig. 1, that correlation is –0.96. A number that expressed only the ordering relation hypothesized, and not its linearity, would be even closer to –1.0.

One cannot, however, reliably infer an individual's current level of well-being from knowledge of his chronic conditions only. There is no a priori reason to expect a one-to-one correspondence between the number of chronic conditions and level of well-being. On an individual basis, in fact, the correlation of W* with the number of chronic conditions is considerably lower (–0.38). The overall correlation, nevertheless, confirms completely the hypothesis of a consistently decreasing average level of well-being with more chronic conditions. This provides strong convergent evidence for the validity of the underlying construct.

Number of Symptoms or Problems. A second prediction from the construct is that persons reporting more symptom/problem complexes will have lower levels of well-being (both W and W*). Figure 2 portrays the mean symptom-specific level of well-being (W*) according to the number of complexes reported and the number of persons reporting each number of complexes. As

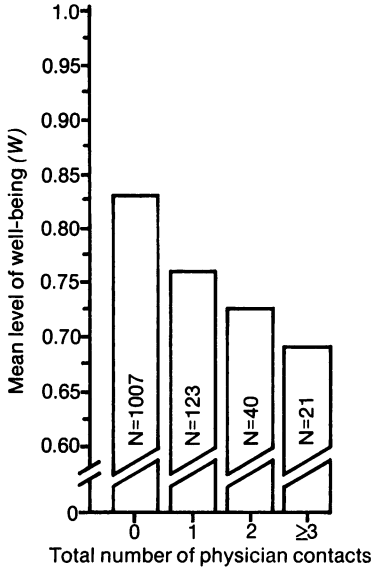

Fig. 3. Mean symptom-standardized level of well-being (W) for groups of persons reporting specific numbers of physician contacts: office visits plus telephone contacts (weighted Pearson's $r = -0.55$).

with chronic conditions, the IWB decreases in an absolutely consistent fashion as the number of complexes increases. The weighted Pearson's r for this relationship is only –0.75, indicating that the relation is not perfectly linear.

The correlation with the reported number of complexes on an individual basis was –0.61 for the symptom-standardized W and –0.62 for the symptom-specific W^*. The relatively high correlation indicates that the index is sensitive to current problems and discomfort, whether from acute or chronic conditions.

As noted with chronic medical conditions, such correlations do not indicate that the number of symptom/problem complexes can substitute for W or W^*. A correlation of –0.62 accounts for only 36 percent of the variance in the individual level of well-being. Nevertheless, the overall correlation again supports the basic construct.

Physician Contacts. We expect that people in lower levels of well-being will use medical services more than persons in higher levels. Figure 3 shows the relation between the mean level of well-being (W) and the total number of physician contacts (visits plus phone calls) in the eight days preceding the 1974 interview. The weighted correlation for this relation is –0.55.

Although attempts were made to exclude them in the interview, telephone calls for appointments and other business matters may inflate the total number

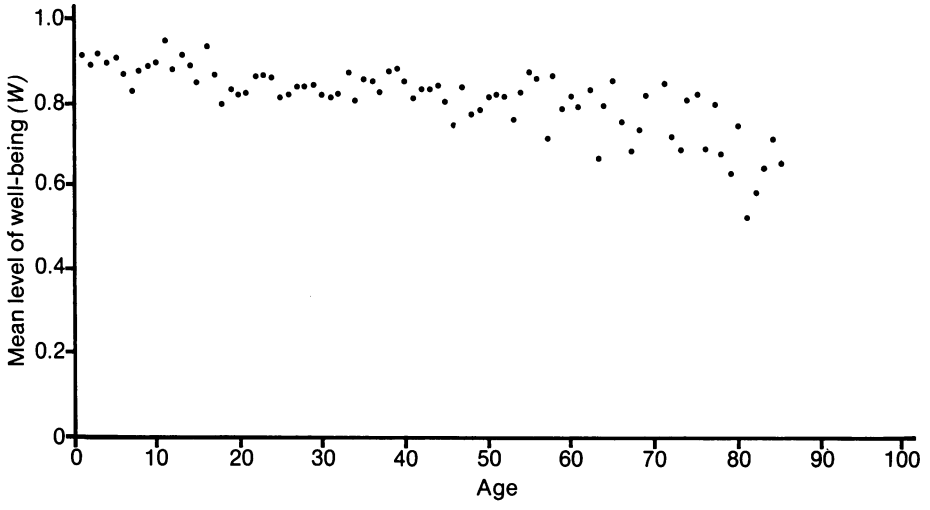

Fig. 4. Three-year moving average symptom-standardized level of well-being (W) for living persons in the age groups shown (weighted Pearson's $r = -0.75$).

of contacts, which would dilute the relationship. That this may have occurred is suggested by the weighted correlation of W with physician visits (excluding calls) of –0.92. In all comparisons, however, the data consistently confirm the expected negative correlation of the index with physician utilization.

Dysfunctional Persons. To increase the precision of our estimates and to exercise our instruments in lower function levels, we used a screening question in the 1974 survey to identify supplementary household members who had not been selected in the probability sample but who had been ill or in some dysfunctional state in the week prior to the interview.

Eighty-nine persons, adults and children, were identified whose level of well-being should logically be significantly less than that of the total sample. The prediction is confirmed with a mean level of well-being (W^*) among dysfunctional persons of 0.63 and a mean W^* of 0.81 among the probability sample of respondents and children ($t = 8.68$, $p < 0.0001$). This provides further convergent evidence of validity.

The precision of the estimates is also worth noting. For the 863 respondents in the sample the standard error of the mean level of well-being (W^*) is 0.005 on a 0–1 scale, and for the 89 dysfunctional persons it is 0.019. Power analyses with this degree of reliability demonstrate that relatively small changes or differences in well-being can be detected with sample sizes that are feasible in household surveys and follow-up studies.

Age. According to our conceptual framework [12], the expected value of the Index of Well-being decreases with greater age for any population. This prediction is based on the well-grounded observation that the older a group of individuals, the lower will be their aggregate function status. Figure 4 shows

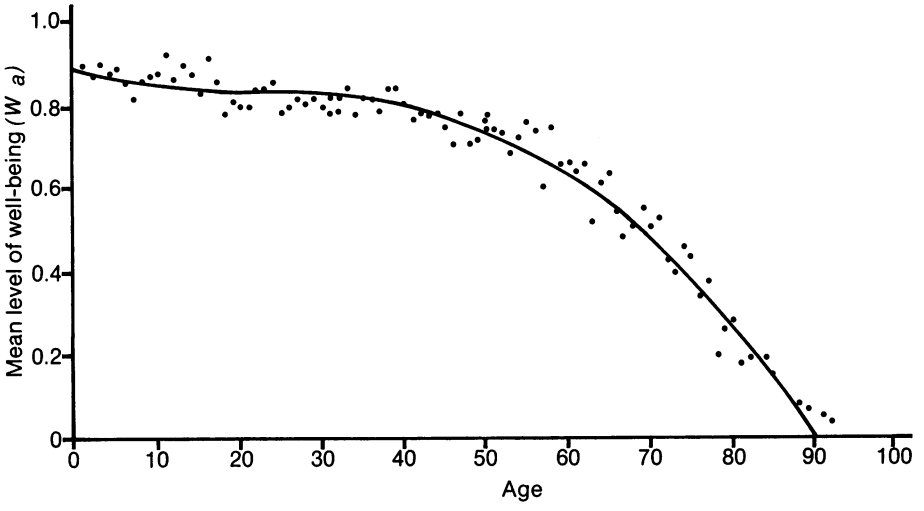

Fig. 5. Mean symptom-standardized level of well-being for age groups, adjusted (W_a) by expected proportion of survivors (from life table) in each age group. The area under the curve approximates *E*, the Weighted Life Expectancy for the population.

the modest decline of the mean symptom-standardized level of well-being, *W*, with increasing age.

Because some of the single years, especially in the older age groups, contained only 15–20 individuals, the points in the figure are three-year moving averages: for example, the point for age 51 in the graph is actually averaged over ages 50–52. Although not dramatic, the decline is very consistent except in the preadolescent period, as expected. When age is correlated with the mean *W* in each specific year, averaged over all non-age-related variation, a Pearson's *r* of –0.75 demonstrates the consistency.

Because it ignores mortality, however, Fig. 4 does not accurately reflect all changes in the level of well-being with age. This is because the points on the graph represent only the survivors of a birth cohort and do not include those who would be the same age but are not now living. Conversely, if everyone in a population were in a completely well state and all deaths occurred suddenly, an index computed on those still living would show no variation at all with age. Only a relation that includes mortality can adequately test the correlation with age of even a time-specific index.

To take proper account of this relation, Fig. 5 displays a practical and powerful extension of the age graph. Its points are found by multiplying *W* for age group *a* by the proportion of persons who would still be living at age *a* according to a life table constructed from current local mortality rates. The area under the curve then represents a simple static or nonstochastic approximation of the Weighted Life Expectancy *E*. A similar and more sensitive approximation can be computed using W^* instead of *W*.

Table 3. Correlations of Self-rated Well-being with W and W* for Each of Eight Days Prior to Interview

Day† prior to interview	Correlation of self-rated well-being	
	With W	With W*
1	0.43	0.46
2	0.43	0.46
3	0.42	0.46
4	0.45	0.48
5	0.45	0.49
6	0.44	0.47
7	0.46	0.48
8	0.42	0.46

† Day 1 is the day before the interview; Day 8 is the eighth day prior.

That approximation goes significantly beyond Sullivan's 1971 proposal [27] for a combined morbidity and mortality index because the measured social preferences integrate *all* differences in dysfunction, no matter how small, into an overall index. The curve in the graph demonstrates the sensitivity of the Index of Well-being to changes in age when mortality is included. The specific polynomial describing this curve is

$$W = 0.8835 - 0.00623\,(\text{AGE}) + 0.000255\,(\text{AGE})^2 - 0.0000047\,(\text{AGE})^3 + 0.000000016\,(\text{AGE})^4$$

All coefficients are statistically significant at the 0.01 level. The sensitivity of this relationship provides strong convergent evidence for the Index of Well-being.

Self-rated Well-being. If the time-specific measure of well-being is valid, then persons with high W and W* values should perceive their health situations for any single day as more desirable, on the average, than the self-perceptions of those on the lower end of the well-being scale. Thus we would predict a *positive* relation between W and self-perceived well-being, where the term well-being is used for a "one day only" rating and the term "health status" is reserved for perceptions of overall health that involve future expectations.

The most sensitive measures of this relationship were obtained in our 1975 reinterview survey. Respondents (and parents for their children) gave direct ratings on a 0–10 scale, where 0 was death and 10 was completely well, for the eight individual days prior to the interview. They also rated their overall health status, "taking the future into account," on a 0–10 scale. The self-rating was done immediately after the training and experience of a 20-minute preference measurement exercise, so the respondents were quite familiar with the procedure. (Our use of the self-rated health status will be described later.)

Table 3 displays the correlation of W and W* with self-rated well-being for each of the eight days prior to the interview. The correlations (0.42 to 0.46 for W and 0.46 to 0.49 for W*) are substantial and in the expected direction. These figures are equivalent to a mean R^2 of 0.192 for W as an explanatory

variable and a mean R^2 of 0.223 for W^*. The difference between the two R^2s (0.0318) is equivalent to a 17-percent increase in the variance in self-rated well-being explained by the symptom-specific W^*. Although small, such increased precision in reflecting self-rated well-being is highly desirable, especially for outcome and other studies which may involve fewer respondents than do surveys.

Using W^* as the better-correlated variable, a more accurate picture of its relation to self-rated well-being is given by taking the mean self-rated well-being of all persons who share a common function level and symptom/problem complex. Since it is not the self-rating of a particular respondent that we wish to measure, but the social rating for all similar cases, the mean of the self-ratings helps reduce the "noise" due to measurement error, interindividual and day-to-day intra-individual variations, sampling error, and confounding interactions.

One hundred twenty-four different combinations of the function status factors (function level and symptom/problem complex, $K = 124$ in Eq. 3) occurred in our 1974 survey among respondents and children, and the weighted correlation of the W^*s with the mean self-ratings of identically classified respondents is 0.76. This indicates substantial success at representing the shared component of respondents' self-ratings with the standardized and highly reliable scoring system that yields W^*.

A more sensitive test of W^* compares it with self-rated well-being for respondents in function level L42—those who reported no functional limitations but had some symptom/problem complex. Most health indicators do not differentiate at all among such persons, who constitute about 50 percent of household survey respondents. From 1974 data we computed the mean self-rated well-being for all such persons who selected the same symptom/problem complex as "most undesirable." The correlation between W^* and the mean self-rated well-being, weighted by the number of respondents with each symptom, was 0.63. This correlation further demonstrates that the index sensitively represents the shared component of subjective self-assessments of well-being, even in the topmost steps (5,4,5) of the function status scales.

Even pooled self-ratings, however, cannot serve as an adequate criterion for the ideal index values, since there is no evidence for a correspondence of self-ratings to the social preferences that are implied by consumers' choices about medical care, either for themselves or as ethical preferences for public policy. On the other hand, evidence for such a direct correspondence between our levels of well-being and the preferences implied by social choices has already been established by rigorous studies [15]. Furthermore, the preferences used in calculating W^* were computed from over 40,000 ratings by the consumer respondents themselves, in a set selected so that precise estimates could be made for *all possible* combinations of function level and symptom/problem complex, and not simply for those cases that happen to occur in a particular study or survey.

The measured preference weights show substantial consistency with self-rated well-being, providing still further convergent evidence for the construct

Table 4. Correlations (Pearson's *r*) Among Daily Self-ratings of Well-being and Daily Computed Values of *W**
(*N* varies from 885 to 891)

Day† of self-rating	Day† for which *W** was computed							
	1	2	3	4	5	6	7	8
1	**0.46**	0.45	0.40	0.38	0.38	0.37	0.37	0.37
2	0.43	**0.46**	0.41	0.39	0.36	0.35	0.35	0.35
3	0.43	0.45	**0.46**	0.45	0.40	0.39	0.38	0.37
4	0.40	0.44	0.44	**0.48**	0.44	0.43	0.42	0.40
5	0.38	0.40	0.40	0.43	**0.49**	0.44	0.42	0.40
6	0.36	0.38	0.37	0.39	0.43	**0.47**	0.46	0.44
7	0.35	0.40	0.36	0.38	0.40	0.45	**0.48**	0.46
8	0.35	0.37	0.36	0.36	0.39	0.43	0.45	**0.47**

† Day 1 is the day before the day of interview; Day 8 is the eighth day prior.

validity of the Index of Well-being. Having demonstrated the consistency of the index in all available comparisons in which unpredicted results would argue against its validity, we now turn to the somewhat more complex assessment of the discriminant evidence for construct validity.

Discriminant Evidence

Discriminant evidence indicates that the measure does not represent a construct other than the one it is devised to measure. That is, it correlates more strongly with measures that are more closely related to the construct than with other measures that bear a looser relation to the construct [23]. Two analyses help establish the discriminant validity of the IWB.

The first comparison correlates *W** for each of the eight days preceding the 1975 interview with self-rated well-being for the same day and with self-ratings for each of the other seven days. The matrix of correlations is shown in Table 4. If the Index of Well-being is really a sensitive time-specific measure, *W** should correlate most highly with self-ratings for the specific day on which *W** is assessed, that is, the correlations on the diagonal of the matrix (shown in boldface) should be higher than the off-diagonal entries.

Table 4 demonstrates not only that the diagonal terms are the largest (r = 0.46 to 0.49) but that the association decreases systematically (to about 0.36) as the time between the different assessments becomes longer. Furthermore, the same-day correlations between self-ratings and *W** from one to eight days ago (along the diagonal) are not attenuated by any effects of memory over the eight days. These data support the notion that the Index of Well-being discriminates among adjacent days and reliably reflects small day-to-day variations in well-being.

The second analysis for discriminant validity correlates *W** both with self-rated well-being on a particular day and with self-rated health status. For self-rated health status, the respondents were trained to include the "outlook for the future" (prognoses), which the IWB specifically excludes; we therefore

predicted that W^* would correlate more highly with self-ratings of current well-being than with the self-ratings of overall health status.

As predicted, the correlation between W^* on day 1 (yesterday) and individual self-rated health status was significantly lower ($r = 0.09$) than the correlations already noted between W^* and self-rated well-being on day 1 ($r = 0.46$). Weighted correlations of W^* with mean self-rated health status would surely amplify the monotonic relation between the two measures by averaging over individual variations to reveal the underlying pattern, but the basic relation between the three measures, which are always required for discriminant validity, would remain the same.

On an individual basis, in fact, the correlation of W^* for one day with self-rated overall health status is so low that W^* appears to give almost no information about expected future well-being as perceived by a single respondent. The marked divergence in the two measures dramatically underscores the fact that consumers recognize the difference between their current level of well-being and their prognostic outlook. This difference provides substantial discriminant evidence for the validity of separating prognoses from the time-specific dimension of well-being in the basic health status construct.

Discussion

Perhaps because its development has been reported step by step and piece by piece in numerous articles over the past several years [1–3,12–20], the Index of Well-being has been seen as complex and difficult to comprehend. Creating a health index that will answer the many legitimate practical and theoretical questions that can be raised is an inherently complex task. The index that results from the research, however, need not be difficult to comprehend or apply. We hope that careful consideration of the IWB, and the Weighted Life Expectancy of which it is a component, will reveal that: the notion of health as an expectation is widely accepted; the separate determination of prognoses is not only necessary but feasible; and the relative desirability of various dysfunctions can be meaningfully and reliably incorporated only by using standardized measures of social preferences—preferences that seem to vary little across many different cultural groups.

In this article we have tried to address the question of validity in the context of the most rigorous formal definitions of that concept known to us. In all tests to date the proposed index fulfills those definitions that are relevant:

- Since no single directly observable measure of well-being exists, testing for criterion validity is inappropriate.
- The proposed index demonstrates content validity by including all possible levels of function and symptom/problem complexes and a clear relation to the death state, as well as consumer ratings of the relative importance of the states.
- Data from a metropolitan household interview survey provide convergent evidence of construct validity by demonstrating an expected positive

correlation of the index with self-rated well-being and expected negative correlations with age, number of chronic medical conditions, number of reported symptoms or problems, number of physician contacts, and dysfunctional status.

- Differences in correlation between current well-being and self-assessed overall health status, and between the symptom-specific well-being and self-rated well-being on different days, exhibit discriminant evidence of construct validity.

We will continue to investigate the validity of the proposed index, both by further analyses of the data on hand and with data yet to be gathered.

The relationship between age and well-being shown in Fig. 5 has significant potential as a social indicator of health. As previously mentioned (p. 500), the simple weighted average level of well-being, adjusted by current local mortality rates, yields a static or nonstochastic approximation of the Weighted (quality-adjusted) Life Expectancy *E*. This comprehensive index combines acute and chronic illness, integrates all levels of dysfunction including symptoms that produce no limitation of activity, avoids completely the paradoxical inflation of health reported by indicators that do not take account of mortality, and yet is computable with data from a single cross-sectional survey using simple arithmetic. It is usable now as a reliable comprehensive social indicator for health, on the national level or in smaller areas. In 1974 in San Diego County, for example, the unweighted life expectancy was 71.9 years. If the mean Index of Well-being for each age group is multiplied by the proportion of persons expected to survive to that age, a synthetic cohort is created, with a Weighted Life Expectancy of 58.6 well-years, the area under the curve in Fig. 5.

The difference between 58.6 well-years and 71.9 expected years represents an average of 13.3 years of life of diminished quality for each resident of San Diego County. It is to this gap—to the quality of life—that health planning, improvements in health care delivery, medical research, preventive medicine, and programs to produce changes in lifestyle should be addressed, perhaps as much as to extensions of the life expectancy itself.

Acknowledgments. The authors thank John P. Anderson, W.R. Blischke, Martin Chen, Daniel Tunstall, Thomas Wan, and John Ware for helpful comments. The authors gratefully acknowledge the collaboration of John Scott, Charles Cannell, and other personnel of the University of Michigan Survey Research Center, which conducted the special 1974 survey.

REFERENCES

1. Bush, J.W., W.R. Blischke, and C.C. Berry. Health Indices, Outcomes, and the Quality of Medical Care. In R. Yaffe and D. Zalkind (eds.), *Evaluation in Health Services Delivery*, pp. 313–339. New York: Engineering Foundation, 1975.
2. Chen, M.M. and J.W. Bush. Maximizing health system output with political and administrative constraints using mathematical programming. *Inquiry* 13:215 Sept. 1976.
3. Chen, M.M., J.W. Bush, and D.L. Patrick. Social indicators for health planning and policy analysis. *Policy Sci* 6:71 Mar. 1975.
4. *Health: United States, 1975.* DHEW Pub. No. (HRA) 76-1232. Washington, DC: U.S. Government Printing Office, 1976.
5. Sullivan, D.F. *Conceptual Problems in Developing an Index of Health.* PHS Pub. No. 1000. Series 2, No. 17. Washington, DC: U.S. Government Printing Office, 1966.

6. Moriyama, I.M. Problems in the Measurement of Health Status. In E. Sheldon and W. Moore (eds.), *Indicators of Social Change*, pp. 573–600. New York: Russell Sage, 1968.
7. Goldsmith, S.B. The status of health status indicators. *Health Serv Rep* 87:212 Mar. 1972.
8. Goldsmith, S.B. A reevaluation of health status indicators. *Health Serv Rep* 88:937 Dec. 1973.
9. Miller, J.E. Guidelines for Selecting a Health Status Index: Suggested Criteria. In R.L. Berg (ed.), *Health Status Indexes*, pp. 243–247. Chicago: Hospital Research and Educational Trust, 1973.
10. Balinsky, W. and R. Berger. A review of the research on general health status indexes. *Med Care* 13:283 Apr. 1975.
11. Torrance, G.W. Health status index models: A unified mathematical view. *Manage Sci* 22:990 May 1976.
12. Fanshel, S. and J.W. Bush. A health status index and its application to health services outcomes. *Oper Res* 18:1021 Nov.–Dec. 1970.
13. Bush, J.W., S. Fanshel, and M.M. Chen. Analysis of a tuberculin testing program using a health status index. *J Socio-Economic Plann Sci* 6:49 Feb. 1972.
14. Patrick, D.L., J.W. Bush, and M.M. Chen. Toward an operational definition of health. *J Health Soc Behav* 14:6 Mar. 1973.
15. Patrick, D.L., J.W. Bush, and M.M. Chen. Methods for measuring levels of well-being for a health status index. *Health Serv Res* 8:228 Fall 1973.
16. Blishcke, W.R., J.W. Bush, and R.M. Kaplan. A successive intervals analysis of social preference measures for a health status index. *Health Serv Res* 10:181 Summer 1975.
17. Bush, J.W., M.M. Chen, D.L. Patrick, and W.R. Blischke. *Statistical Models of Social Preferences for Constructing a Health Status Index.* Pub. No. 236 155/8ST. Springfield, VA: National Technical Information Service, 1974.
18. Kaplan, R.M. and J.W. Bush. Comparison of magnitude estimation and category rating for measuring preference in a health status index. Working paper, Division of Health Policy, University of California at San Diego, 1976.
19. Kaplan, R.M. and J.W. Bush. Multitrait-multimethod comparison of preference measurements for a health status index. Paper presented at the Western Psychological Association meeting, Sacramento, CA, Mar. 1975.
20. Bush, J.W., M.M. Chen, and J. Zaremba. Estimating health program outcomes using a Markov equilibrium analysis of disease development. *Am J Public Health* 61:2362 Dec. 1971.
21. Cronbach, L.J. Test Validation. In R.L. Thorndike (ed.), *Educational Measurement*, 2nd ed., pp. 443–507. Washington, DC: American Council on Education, 1971.
22. American Psychological Association. *Standards for Educational and Psychological Tests.* Washington, DC: American Psychological Association, 1974.
23. Campbell, D.T. and D.W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 56:81 Mar. 1959.
24. Nunnally, J.C. *Psychometric Theory.* New York: McGraw-Hill, 1967.
25. Kuhn, T.S. *The Structure of Scientific Revolutions*, 2nd ed. Chicago: University of Chicago Press, 1970.
26. Reynolds, W.J., W.A. Rushing, and D.L. Miles. The validation of a Function Status Index. *J Health Soc Behav* 15:271 Dec. 1974.
27. Sullivan, D.F. A single index of mortality and morbidity. *HSMHA Health Rep* 86:347 Apr. 1971.
28. Cronbach, L.J. and P.E. Meehl. Construct validity in psychological tests. *Psychol Bull* 52:281 July 1955.
29. Glasser, J.H. and R.N. Forthofer. Analysis of health service data. Department of Biometry, University of Texas at Houston, 1972.
30. Kogan, L.S. and S. Jenkins. *Indicators of Child Health and Welfare: Development of the DIPOV Index.* New York: Columbia University Press, 1974.
31. Hively, W., H.L. Patterson, and S.H. Page. A "universe-defined" system of arithmetic achievement tests. *J Educ Meas* 5:275 Winter 1968.
32. Stewart, A., J.E. Ware, and R.H. Brook. The meaning of health: Understanding functional limitations. Paper presented to the Statistics Section, American Public Health Association meeting, Miami, FL, Oct. 1976.

33. Berg, R.L., D.S. Hallauer, and S.N. Berk. Neglected aspects of the quality of life. *Health Serv Res* 11:391 Winter 1976.
34. Lord, F.M. and M.R. Novick. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley, 1968.
35. Smith, D.B. and A. Kaluzny. *The White Labyrinth: Understanding the Organization of Health Care.* Berkeley, CA: McCutchan, 1975.
36. Raiffa, H. *Decision Analysis.* Reading, MA: Addison-Wesley, 1968.
37. Torrance, G.W. Social preference for health states: An empirical evaluation of three measurement techniques. *Socio-Economic Plann Sci* 10(3):129, 1976.
38. Torrance, G.W. Toward a utility theory foundation for health status index models. *Health Serv Res* 11:349 Winter 1976.
39. Chiang, C.L. and R.D. Cohen. How to measure health: A stochastic model for an index of health. *Int J Epidemiol* 2:7 Spring 1973.
40. Parsons, T. *The Social System.* New York: The Free Press, 1951.
41. Fabrega, H. Jr. The need for an ethnomedical science. *Science* 189:969 Sept. 9, 1975.
42. Anderson, N.H. Information Integration Theory: A Brief Survey. In D.H. Krantz, R.C. Atkinson, R.L. Luce, and P. Suppes (eds.), *Contemporary Developments in Mathematical Psychology,* Vol. 2, pp. 236–305. San Francisco: Freeman, 1974.
43. Grossman, M. *The Demand for Health: A Theoretical and Empirical Investigation.* New York: National Bureau of Economic Research, 1972.
44. Grossman, M. On the concept of health capital and the demand for health. *J Polit Econ* 80:223 Apr. 1972.
45. Messick, S. *The Standard Problem: Meaning and Values in Measurement and Evaluation.* Educational Testing Service Research Bulletin RB-74-44. Princeton, NJ: Educational Testing Service, Oct. 1974.